e-Appendix B

Linear Algebra

The basic storage structure for the data is a matrix X which has N rows, one for each data point; each data point is a row-vector.

$$\mathrm{X} = \begin{bmatrix} \text{—}\mathbf{x}_1^{\mathrm{T}}\text{—} \\ \text{—}\mathbf{x}_2^{\mathrm{T}}\text{—} \\ \vdots \\ \text{—}\mathbf{x}_N^{\mathrm{T}}\text{—} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ & \vdots & & \\ x_{N,1} & x_{N,2} & \cdots & x_{N,d} \end{bmatrix}$$

When we write $\mathrm{X} \in \mathbb{R}^{N\times d}$ we mean a matrix like the one above, which in this case is $N \times d$ (N rows and d columns). Linear algebra plays an important role in manipulating such matrices when learning from data, because the matrix X can be viewed as a linear operator that takes a set of weights $\mathbf{w}$ and outputs in-sample predictions:

$$\hat{\mathbf{y}} = \mathrm{X}\mathbf{w}.$$

B.1 Basic Properties of Vectors and Matrices

A (column) vector $\mathbf{v} \in \mathbb{R}^d$ has d components $v_1, \ldots, v_d$. The matrix X above could be viewed as a set of d column vectors or a set of N row vectors. A vector can be multiplied by a scalar in the usual way, by multiplying each component by the scalar, and two vectors can be added together by adding the respective components together.

The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^d$ are linearly independent if no non-trivial linear combination of them can equal $\mathbf{0}$:

$$\sum_{i=1}^{m} \alpha_i \mathbf{v}_i = \mathbf{0} \implies \alpha_i = 0 \text{ for } i = 1, \ldots, m.$$

If the vectors are not linearly independent, then they are linearly dependent. Given two vectors $\mathbf{v}, \mathbf{u}$, the standard Euclidean inner product (dot product)

is $\mathbf{v}^{\mathrm{T}}\mathbf{u} = \sum_{i=1}^{d} v_i u_i$ and the Euclidean norm is $\|\mathbf{v}\|^2 = \mathbf{v}^{\mathrm{T}}\mathbf{v} = \sum_{i=1}^{d} v_i^2$. Let θ be the angle between $\mathbf{v}$ and $\mathbf{u}$ in the standard geometric sense. Then

$$\mathbf{v}^{\mathrm{T}}\mathbf{u} = \|\mathbf{v}\|\|\mathbf{u}\|\cos\theta \implies (\mathbf{v}^{\mathrm{T}}\mathbf{u})^2 \le \|\mathbf{v}\|^2\|\mathbf{u}\|^2,$$

where the latter inequality is known as the Cauchy-Schwarz inequality. The two vectors $\mathbf{v}$ and $\mathbf{u}$ are orthogonal if $\mathbf{v}^{\mathrm{T}}\mathbf{u} = 0$; if in addition $\|\mathbf{v}\| = \|\mathbf{u}\| = 1$, then the two vectors are orthonormal (orthogonal and have unit norm).

A basis $\mathbf{v}_1, \ldots, \mathbf{v}_d$ has the property that any $\mathbf{u} \in \mathbb{R}^d$ can be written as a *unique* linear combination of the basis vectors, $\mathbf{u} = \sum_{i=1}^{d} \alpha_i \mathbf{v}_i$. It follows that the basis must be linearly independent. (We have implicitly assumed that the cardinality of any basis is d, which is indeed the case.) A basis $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is an orthonormal basis if $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \delta_{ij}$ (equal to 1 if $i = j$ and zero otherwise), that is the basis vectors are pairwise orthonormal.

Exercise B.1

This exercise introduces some fundamental properties of vectors and bases.

(a) Are the following sets of vectors dependent or independent?

$$\left\{\begin{bmatrix}1\\2\end{bmatrix}\right\} \quad \left\{\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\4\end{bmatrix}\right\} \quad \left\{\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\} \quad \left\{\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}3\\1\end{bmatrix}\right\}$$

(b) Which of the sets in (a) span $\mathbb{R}^2$?

(c) Show that the expansion $\mathbf{u} = \sum_{i=1}^{d} \alpha_i \mathbf{v}_i$ holds for unique α_i if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are independent.

(d) Which of the sets in (a) are a basis for $\mathbb{R}^2$?

(e) Show that any set of vectors containing the zero vector is dependent.

(f) Show: if $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^d$ are independent then $m \le d$ (the maximum cardinality of an independent set is the dimension d). *[Hint: Induction on d.]*

(g) Show: if $\mathbf{v}_1, \ldots, \mathbf{v}_m$ span $\mathbb{R}^d$, then $m \ge d$. *[Hint: The span does not change if you add a multiple of one of the vectors to another. Hence, transform the set to a more convenient one.]*

(h) Show: every basis has cardinality equal to the dimension d.

(i) Show: If $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal, they are independent. If $\mathbf{v}_1, \mathbf{v}_2$ are independent, then $\mathbf{v}_1, \mathbf{v}_2 - \lambda\mathbf{v}_1$ have the same span and are orthogonal, where $\lambda = \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 / \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1$. What if $\mathbf{v}_1, \mathbf{v}_2$ are dependent?

(j) Show that any set of independent vectors can be transformed to a set of pairwise orthogonal vectors with the same span.

(k) Given a basis $\mathbf{v}_1, \ldots, \mathbf{v}_d$ show how to construct an orthonormal basis $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_d$. *[Hint: Start by making $\mathbf{v}_2, \ldots, \mathbf{v}_d$ orthogonal to $\mathbf{v}_1$.]*

(l) If $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is an orthonormal basis, then show that any vector $\mathbf{u}$ has the (unique) expansion $\mathbf{u} = \sum_{i=1}^{d} (\mathbf{u}^{\mathrm{T}}\mathbf{v}_i)\mathbf{v}_i$.

(m) Show: Any set of d linearly independent vectors is a basis for $\mathbb{R}^d$.

If $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is an orthonormal basis, then the coefficients $(\mathbf{u}^{\mathrm{T}}\mathbf{v}_i)$ in the expansion $\mathbf{u} = \sum_{i=1}^{d}(\mathbf{u}^{\mathrm{T}}\mathbf{v}_i)\mathbf{v}_i$ are the coordinates of $\mathbf{u}$ with respect to the (ordered) basis $\mathbf{v}_1, \ldots, \mathbf{v}_d$. These d coordinates form a vector in d dimensions. When we write $\mathbf{u}$ as a vector of its coordinates, as we have been doing, we are implicitly assuming that these coordinates are with respect to the standard orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_d$:

$$\mathbf{e}_1 = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix}0\\0\\1\\\vdots\\0\end{bmatrix}, \quad \ldots, \quad \mathbf{e}_d = \begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix}$$

Matrices. Let $\mathrm{A} \in \mathbb{R}^{N\times d}$ ($N \times d$ matrix) and $\mathrm{B}, \mathrm{C} \in \mathbb{R}^{d\times M}$ be arbitrary real valued matrices. For the (i, j)-th entry of a matrix, we may write A_{ij} or $[\mathrm{A}]_{ij}$ (the latter when we want to explicitly identify A as a matrix). The transpose $\mathrm{A}^{\mathrm{T}} \in \mathbb{R}^{d\times n}$ is a matrix whose entries are given by $[\mathrm{A}^{\mathrm{T}}]_{ij} = [\mathrm{A}]_{ji}$. The matrix-vector and matrix-matrix products are defined in the usual way,

$$\begin{aligned} [\mathrm{A}\mathbf{x}]_i &= \textstyle\sum_{j=1}^{d} A_{ij}x_j, &&\text{where } \mathbf{x} \in \mathbb{R}^d,\ \mathrm{A} \in \mathbb{R}^{N\times d} \text{ and } \mathrm{A}\mathbf{x} \in \mathbb{R}^N; \\ [\mathrm{AB}]_{ij} &= \textstyle\sum_{k=1}^{d} \mathrm{A}_{ik}\mathrm{B}_{kj}, &&\text{where } \mathrm{A} \in \mathbb{R}^{N\times d},\ B \in \mathbb{R}^{d\times M} \text{ and } \mathrm{AB} \in \mathbb{R}^{N\times M}. \end{aligned}$$

In general, when we refer to products of matrices below, assume that all the products exist. Note that $\mathrm{A}(\mathrm{B}+\mathrm{C}) = \mathrm{AB}+\mathrm{AC}$ and $(\mathrm{AB})^{\mathrm{T}} = \mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$. The $d\times d$ identity matrix I_d is the matrix whose columns are the standard basis, having diagonal entries 1 and zeros elsewhere. If A is a square matrix ($d = N$), the inverse A^{-1} (if it exists) satisfies

$$\mathrm{AA}^{-1} = \mathrm{A}^{-1}\mathrm{A} = \mathrm{I}_d.$$

Note that

$$\begin{aligned} (\mathrm{A}^{\mathrm{T}})^{-1} &= (\mathrm{A}^{-1})^{\mathrm{T}}; \\ (\mathrm{AB})^{-1} &= (\mathrm{B})^{-1}\mathrm{A}^{-1}. \end{aligned}$$

A matrix is invertible if and only if its columns (also rows) are linearly independent (in which case the columns are a basis).

If the matrix A has orthonormal columns, then $\mathrm{A}^{\mathrm{T}}\mathrm{A} = \mathrm{I}_d$; if in addition A is square, then it is orthogonal (and invertible). It is often convenient to refer to the columns of a matrix, and we will write $\mathrm{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_d]$. Using this notation, one can write a matrix vector product as

$$\mathrm{A}\mathbf{x} = \sum_{i=1}^{d} x_i\mathbf{a}_i,$$

from which we see that a matrix can be viewed as an operator that takes the input $\mathbf{x}$ and transforms it into a linear combination of its columns. The range of a matrix A is the subspace spanned by its columns, range(A) = span($\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$). It is also useful to define the subspace spanned by the rows of A, which is the range of A^{T}. The dimension of the range of A is called the column-rank of A. Similarly, the dimension of the subspace spanned by the rows is the row-rank of A. It is a useful fact that the row and column ranks are equal, and so we can define the rank of a matrix, denoted ρ, as the dimension of its range. Note that

$$\rho(\mathrm{A}) = \mathrm{rank}(\mathrm{A}) = \mathrm{rank}(\mathrm{A}^{\mathrm{T}}\mathrm{A}) = \mathrm{rank}(\mathrm{A}\mathrm{A}^{\mathrm{T}}).$$

The matrix $\mathrm{A} \in \mathbb{R}^{n\times d}$ has full column rank if rank(A) $= d$; it has full row rank if rank(A) $= N$. Note that rank(A) $\leq \min(N, d)$, since the dimension of a space spanned by ℓ vectors is at most ℓ (see Exercise B.1(g)).

Exercise B.2

Let $\mathrm{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$, $\mathrm{B} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 0 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

(a) Compute (if the quantities exist): AB; A$\mathbf{x}$; B$\mathbf{x}$; BA; $\mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$; $\mathbf{x}^{\mathrm{T}}\mathrm{A}\mathbf{x}$; $\mathrm{B}^{\mathrm{T}}\mathrm{AB}$; A^{-1}.

(b) Show that $\mathbf{x}^{\mathrm{T}}\mathrm{A}\mathbf{x} = \sum_{i=1}^{3}\sum_{j=1}^{3} x_i x_j \mathrm{A}_{ij}$.

(c) Find a $\mathbf{v}$ for which $\mathrm{A}\mathbf{v} = \lambda\mathbf{v}$. What are the possible choices of λ?

(d) How many linearly independent rows are there in B? How many linearly independent columns are there in B?

(e) Given any matrix $\mathrm{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_d]$, let $\mathbf{c}_1, \ldots, \mathbf{c}_r$ be a basis for the span of $\mathbf{a}_1, \ldots, \mathbf{a}_d$. Let C be the matrix whose columns are this basis, $C = [\mathbf{c}_1, \ldots, \mathbf{c}_r]$. Show that for some matrix R, one can write

$$\mathrm{A} = \mathrm{CR}.$$

(i) What is the column-rank of A?

(ii) What are the dimensions of R?

(iii) Show that every row in A is a linear combination of rows in R.

(iv) Hence, show that the dimension of the subspace spanned by the rows of A is at most r, the column-rank. That is,

$$\text{column-rank(A)} \geq \text{row-rank(A)}.$$

(v) Show that column-rank(A) $\leq$ row-rank(A), and hence that the column-rank equals the row-rank. *[Hint: Consider* A^{T}*.]*

B.2 SVD and Pseudo-Inverse

The singular value decomposition (SVD) of a matrix A is one of the most useful matrix decompositions. For the matrix A, assume $d \le N$ and the rank $\rho \le d$. The SVD of A factorizes A into the product of three special matrices:

$$\mathrm{A} = \mathrm{U}\Gamma\mathrm{V}^{\mathrm{T}}.$$

The matrix $\mathrm{U} \in \mathbb{R}^{N\times\rho}$ has orthonormal columns that are called the left singular vectors of A; the matrix $\mathrm{V} \in \mathbb{R}^{d\times\rho}$ has orthonormal columns that are called right singular vectors of A. So,

$$\mathrm{U}^{\mathrm{T}}\mathrm{U} = \mathrm{V}^{\mathrm{T}}\mathrm{V} = \mathrm{I}_\rho.$$

The matrix $\Gamma \in \mathbb{R}^{\rho\times\rho}$ is diagonal and has as its diagonal entries the (positive) singular values of A, $\gamma_i = [\Gamma]_{ii}$. Typically, the singular values are ordered, so that $\gamma_1 \ge \cdots \ge \gamma_\rho$; in this case, the first column of U is called the top left singular vector, and similarly the first column of V is called the top right singular vector. The condition number of A is the ratio of the largest to smallest singular values: $\kappa = \gamma_1/\gamma_\rho$, which plays an important role in the stability of algorithms involving matrices, such as solving the linear regression problem or inverting the matrix. Algorithms exist to compute the SVD in $O(Nd\min(N,d))$ time. If only a few of the top singular vectors and singular values are needed, these can be obtained more efficiently using iterative subspace methods such as power iteration.

If A is not invertible (for example, not square), it is useful to define the Moore-Penrose pseudo-inverse $\mathrm{A}^\dagger$, which satisfies four properties:

$$(i)\ \mathrm{A}\mathrm{A}^\dagger\mathrm{A} = \mathrm{A};\ (ii)\ \mathrm{A}^\dagger\mathrm{A}\mathrm{A}^\dagger = \mathrm{A}^\dagger;\ (iii)\ (\mathrm{A}\mathrm{A}^\dagger)^{\mathrm{T}} = \mathrm{A}\mathrm{A}^\dagger;\ (iv)\ (\mathrm{A}^\dagger\mathrm{A})^{\mathrm{T}} = \mathrm{A}^\dagger\mathrm{A}.$$

The pseudo-inverse functions as an inverse and plays an important role in linear regression.

Exercise B.3

If the SVD of A is $\mathrm{A} = \mathrm{U}\Gamma\mathrm{V}^{\mathrm{T}}$, by checking all four properties, verify that

$$\mathrm{A}^\dagger = \mathrm{V}\Gamma^{-1}\mathrm{U}^{\mathrm{T}}$$

is a pseudo-inverse of A. Also show that $(\mathrm{A}^{\mathrm{T}})^\dagger = (\mathrm{A}^\dagger)^{\mathrm{T}}$.

If A and B both have rank d or if one of $\mathrm{A}, \mathrm{B}^{\mathrm{T}}$ has orthonormal columns, then $(\mathrm{AB})^\dagger = (\mathrm{B})^\dagger\mathrm{A}^\dagger$. The matrix $\Pi_{\mathrm{A}} = \mathrm{A}\mathrm{A}^\dagger = \mathrm{U}\mathrm{U}^{\mathrm{T}}$ is the projection operator onto the range (column-space) of A; this projection operator can be used to decompose a vector $\mathbf{x}$ into two orthogonal parts, the part $\mathbf{x}_{\mathrm{A}}$ in the range of A and the part $\mathbf{x}_{\mathrm{A}_\perp}$ orthogonal to the range of A:

$$\mathbf{x} = \underbrace{\Pi_{\mathrm{A}}\mathbf{x}}_{\mathbf{x}_{\mathrm{A}}} + \underbrace{(\mathrm{I}-\Pi_{\mathrm{A}})\mathbf{x}}_{\mathbf{x}_{\mathrm{A}_\perp}}.$$

A projection operator satisfies $\Pi^2 = \Pi$. For example $\Pi_{\mathrm{A}}^2 = (\mathrm{A}\mathrm{A}^\dagger\mathrm{A})\mathrm{A}^\dagger = \mathrm{A}\mathrm{A}^\dagger$.

B.3 Symmetric Matrices

A square matrix is symmetric if $A = A^{\mathrm{T}}$ (anti-symmetric if $A = -A^{\mathrm{T}}$). Symmetric matrices play a special role because the covariance matrix of the data is symmetric. If X is the data matrix, then the centered data matrix X_{cen} is obtained by subtracting from each data point the mean vector

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}X^{\mathrm{T}}\mathbf{1},$$

where $\mathbf{1}$ is the N-dimensional vector of ones. In matrix form,

$$\begin{aligned} X_{\text{cen}} &= X - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} \\ &= X - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathrm{T}}X \\ &= \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)X. \end{aligned}$$

From this expression, it is easy to see why $\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)$ is called the centering operator. The covariance matrix Σ is given by

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} = \frac{1}{N}X_{\text{cen}}^{\mathrm{T}}X_{\text{cen}}.$$

Via the decomposition $A = \frac{1}{2}(A + A^{\mathrm{T}}) + \frac{1}{2}(A - A^{\mathrm{T}})$, every matrix can be decomposed into the sum of its symmetric part and its anti-symmetric part. Analogous to the SVD, the spectral theorem says that any symmetric matrix admits a spectral or eigen-decomposition of the form

$$A = U\Lambda U^{\mathrm{T}},$$

where U has orthonormal columns that are the *eigenvectors* of A, so $U^{\mathrm{T}}U = I_\rho$ and Λ is diagonal with entries $\lambda_i = [\Lambda]_{ii}$ that are the eigenvalues of A. Each column $\mathbf{u}_i$ of U for $i = 1, \ldots, \rho$ is an eigenvector of A with eigenvalue λ_i (the number of non-zero eigenvalues is equal to the rank). Via the identity $AU = U\Lambda$, one can verify that the eigenvalues and corresponding eigenvectors satisfy the relation

$$A\mathbf{u}_i = \lambda_i\mathbf{u}_i.$$

All eigenvalues of a symmetric matrix are real. Eigenvalues and eigenvectors can be defined more generally using the above equation (even for non-symmetric matrices), but we will only be concerned with symmetric matrices. If the λ_i are ordered, with $\lambda_1 \geq \cdots \geq \lambda_\rho$, then λ_1 is the top eigenvalue, with corresponding eigenvector $\mathbf{u}_1$, and so on. Note that the eigen-decomposition is similar to the SVD ($A = U\Gamma V^{\mathrm{T}}$) with the flexibility to have negative entries along the diagonal in Λ. Since $AA^{\mathrm{T}} = U\Lambda^2U^{\mathrm{T}} = U\Gamma^2U^{\mathrm{T}}$, one identifies that $\lambda_i^2 = \gamma_i^2$. That is, up to a sign, the eigenvalues and singular values of a symmetric matrix are the same.

B.3.1 Positive Semi-Definite Matrices

The matrix A is symmetric positive semi-definite (SPSD) if it is symmetric and for every non-zero $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^{\mathrm{T}} \mathrm{A} \mathbf{x} \geq 0.$$

One writes $\mathrm{A} \succeq 0$; if for every non-zero $\mathbf{x}$,

$$\mathbf{x}^{\mathrm{T}} \mathrm{A} \mathbf{x} > 0,$$

then A is positive definite (PD) and one writes $\mathrm{A} \succ 0$. If A is positive definite, then $\lambda_i > 0$ (all its eigenvalues are positive), in which case $\lambda_i = \gamma_i$ and the eigen-decomposition and the SVD are identical. We can write $\Lambda = \mathrm{S}^2$ and identify $\mathrm{A} = \mathrm{US}(\mathrm{US})^{\mathrm{T}}$ from which one deduces that every SPSD matrix has the form $\mathrm{A} = \mathrm{ZZ}^{\mathrm{T}}$. The covariance matrix is an example of an SPSD matrix.

B.4 Trace and Determinant

The trace and determinant are defined for a square matrix A (so $N = d$). The trace is the sum of the diagonal elements,

$$\text{trace}(\mathrm{A}) = \sum_{i=1}^{N} \mathrm{A}_{ii}.$$

The trace operator is cyclic:

$$\text{trace}(\mathrm{AB}) = \text{trace}(\mathrm{BA})$$

(when both products are defined). From the cyclic property, if O has orthonormal columns (so $\mathrm{O}^{\mathrm{T}}\mathrm{O} = \mathrm{I}$), then $\text{trace}(\mathrm{OAO}^{\mathrm{T}}) = \text{trace}(\mathrm{A})$. Setting $\mathrm{O} = \mathrm{U}$ from the eigen-decomposition of A,

$$\text{trace}(\mathrm{A}) = \text{trace}(\Lambda) = \sum_{i=1}^{N} \lambda_i$$

(sum of eigenvalues). Note that $\text{trace}(\mathrm{I}_k) = k$.

The determinant of A, written $|\mathrm{A}|$, is most easily defined using the totally anti-symmetric Levi-Civita symbol $\varepsilon_{i_1,\ldots,i_N}$ (if you swap any two indices then the value negates, and $\varepsilon_{1,2,\ldots,N} = 1$; if any index is repeated, the value is zero):

$$|\mathrm{A}| = \sum_{i_1,\ldots,i_N=1}^{N} \varepsilon_{i_1,\ldots,i_N} \mathrm{A}_{1i_1} \cdots \mathrm{A}_{Ni_N}.$$

Since every summand has exactly one term from each column, if you scale any column, the determinant is correspondingly scaled. The anti-symmetry

of $\varepsilon_{i_1,\ldots,i_N}$ implies that if you swap any two columns (or rows) of A, then the determinant changes sign. If any two columns are identical, by swapping them, the determinant must change sign, and hence the determinant must be zero. If you add a vector to a column, then the determinant becomes a sum,

$$|[\mathbf{a}_1,\ldots,\mathbf{a}_i+\mathbf{v},\ldots,\mathbf{a}_N]| = |[\mathbf{a}_1,\ldots,\mathbf{a}_i,\ldots,\mathbf{a}_N]| + |[\mathbf{a}_1,\ldots,\mathbf{v},\ldots,\mathbf{a}_N]|.$$

If $\mathbf{v}$ is a multiple of a column then the second term vanishes, and so adding a multiple of one column to another column does not change the determinant. By separating out the summation with respect to i_1 in the definition of the determinant (the first row of A), we get a useful recursive formula for the determinant known as its expansion by cofactors. We illustrate for a 3×3 matrix below:

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = A_{11}\begin{vmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{vmatrix} - A_{12}\begin{vmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{vmatrix} + A_{13}\begin{vmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{vmatrix}$$

$$= A_{11}(A_{22}A_{33} - A_{23}A_{32}) - A_{12}(A_{21}A_{33} - A_{23}A_{31}) + A_{13}(A_{21}A_{32} - A_{22}A_{31}).$$

(Notice the alternating signs as we expand along the first row.) Geometrically, the determinant of A equals the volume enclosed by the parallelepiped whose sides are the column vectors $\mathbf{a}_1,\ldots,\mathbf{a}_N$. This geometric fact is important when changing variables in a multi-dimensional integral, because the infinitesimal volume element transforms according to the determinant of the Jacobian. A useful fact about the determinant of a product of square matrices is

$$|AB| = |A||B|.$$

Exercise B.4

Show the following useful properties of determinants.

(a) $|I_k| = 1$; if D is diagonal, then $|D| = \prod_{i=1}^{N}[D]_{ii}$.

(b) Show that $|A| = \dfrac{1}{N!}\displaystyle\sum_{i_1,\ldots,i_N}\sum_{j_1,\ldots,j_N}\varepsilon_{i_1\ldots i_N}\varepsilon_{j_1\ldots j_N}A_{i_1j_1}\cdots A_{i_Nj_N}$.

(c) Using (b), argue that $|A| = |A^T|$.

(d) If O is orthogonal (square with orthonormal columns, i.e. $O^TO = I_N$), then $|O| = \pm 1$ and $|OAO^T| = |A|$. *[Hint: $|O^TO| = |O||O^T|$.]*

(e) $|A^{-1}| = 1/|A|$. *[Hint: $|A^{-1}A| = |I_n|$.]*

(f) For symmetric A,

$$|A| = \prod_{i=1}^{n}\lambda_i \qquad \text{(product of eigenvalues of A)}.$$

[Hint: $A = U\Lambda U^T$, where U is orthogonal.]

B.4.1 Inverse and Determinant Identities

The inverse of the covariance matrix plays a role in many learning algorithms. Hence, it is important to be able to update the inverse efficiently for small changes in a matrix. If A and B are square and invertible, then the following useful identities are known as the Sherman-Morrison-Woodbury inversion and determinant formulae:

$$\begin{aligned}(\mathrm{A}+\mathrm{XBY}^{\mathrm{T}})^{-1} &= \mathrm{A}^{-1}-\mathrm{A}^{-1}\mathrm{X}(\mathrm{B}^{-1}+\mathrm{Y}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{X})^{-1}\mathrm{Y}^{\mathrm{T}}\mathrm{A}^{-1};\\ |\mathrm{A}+\mathrm{XBY}^{\mathrm{T}}| &= |\mathrm{A}||\mathrm{B}||\mathrm{B}^{-1}+\mathrm{Y}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{X}|.\end{aligned}$$

An important special case is the inverse and determinant updates due to a rank 1 update of a matrix. Setting $\mathrm{X}=\mathbf{x}$, $\mathrm{B}=1$ and $\mathrm{Y}=\mathbf{y}$,

$$\begin{aligned}(\mathrm{A}+\mathbf{x}\mathbf{y}^{\mathrm{T}})^{-1} &= \mathrm{A}^{-1}-\frac{\mathrm{A}^{-1}\mathbf{x}\mathbf{y}^{\mathrm{T}}\mathrm{A}^{-1}}{1+\mathbf{y}^{\mathrm{T}}\mathrm{A}^{-1}\mathbf{x}};\\ |\mathrm{A}+\mathbf{x}\mathbf{y}^{\mathrm{T}}| &= |\mathrm{A}|(1+\mathbf{y}^{\mathrm{T}}\mathrm{A}^{-1}\mathbf{x}).\end{aligned}$$

Setting $\mathbf{x}=\pm\mathbf{y}$ gives the updates for $\mathrm{A}\pm\mathbf{x}\mathbf{x}^{\mathrm{T}}$. If A has a block representation, we can get similar updates to the inverse and determinant. Let

$$\mathrm{A}=\begin{bmatrix}\mathrm{A}_{11} & \mathrm{A}_{12}\\ \mathrm{A}_{21} & \mathrm{A}_{22}\end{bmatrix},$$

and define

$$\begin{aligned}\mathrm{F}_1 &= \mathrm{A}_{11}-\mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21}\\ \mathrm{F}_2 &= \mathrm{A}_{22}-\mathrm{A}_{21}\mathrm{A}_{11}^{-1}\mathrm{A}_{12}\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{A}^{-1} &= \begin{bmatrix}\mathrm{F}_1^{-1} & -\mathrm{A}_{11}^{-1}\mathrm{A}_{12}\mathrm{F}_2^{-1}\\ -\mathrm{F}_2^{-1}\mathrm{A}_{21}\mathrm{A}_{11}^{-1} & \mathrm{F}_2^{-1}\end{bmatrix}, \qquad \text{and}\\ |\mathrm{A}| &= |A_{22}||\mathrm{F}_1| = |\mathrm{A}_{11}||\mathrm{F}_2|.\end{aligned}$$

Again, an important special case is when

$$\mathrm{A}=\begin{bmatrix}\mathrm{X} & \mathbf{b}\\ \mathbf{b}^{\mathrm{T}} & c\end{bmatrix}.$$

In this case,

$$\begin{aligned}\mathrm{A}^{-1} &= \begin{bmatrix}\mathrm{X}^{-1} & \mathbf{0}\\ \mathbf{0}^{\mathrm{T}} & 0\end{bmatrix}+\frac{1}{c-\mathbf{b}^{\mathrm{T}}\mathrm{X}^{-1}\mathbf{b}}\begin{bmatrix}\mathrm{X}^{-1}\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathrm{X}^{-1} & -\mathrm{X}^{-1}\mathbf{b}\\ -\mathbf{b}^{\mathrm{T}}\mathrm{X}^{-1} & 1\end{bmatrix}, \qquad \text{and}\\ |\mathrm{A}| &= |\mathrm{X}|(c-\mathbf{b}^{\mathrm{T}}\mathrm{X}^{-1}\mathbf{b}).\end{aligned}$$

Exercise B.5

Use the formula for the determinant of a 2×2 block matrix to show Sylvester's determinant theorem for matrices $\mathrm{A} \in \mathbb{R}^{n\times d}$ and $\mathrm{B} \in \mathbb{R}^{d\times n}$:

$$|\mathrm{I}_n + \mathrm{AB}| = |\mathrm{I}_d + \mathrm{BA}|.$$

$\left[\textit{Hint: Consider } \begin{vmatrix} \mathrm{I}_n & \mathrm{A} \\ -\mathrm{B} & \mathrm{I}_d \end{vmatrix}\right]$

Use Sylvester's theorem to show the Sherman-Morrison-Woodbury determinant identity

$$|\mathrm{A} + \mathrm{XBY}^{\mathrm{T}}| = |\mathrm{A}||\mathrm{B}||\mathrm{B}^{-1} + \mathrm{Y}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{X}|.$$

[Hint: $\mathrm{A} + \mathrm{XBY}^{\mathrm{T}} = \mathrm{A}(\mathrm{I} + \mathrm{A}^{-1}\mathrm{XBY}^{\mathrm{T}})$*, and use Sylvester's theorem.]*

B.5 Inner Products, Matrix and Vector Norms

For vectors $\mathbf{x}$ and $\mathbf{z}$, the inner product is a symmetric, bilinear positive definite function usually denoted $\langle \mathbf{x}, \mathbf{z} \rangle$. The standard Euclidean inner product (dot product) is and example:

$$\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x} \bullet \mathbf{z} = \mathbf{x}^{\mathrm{T}}\mathbf{z} = \sum_{i=1}^{d} x_i z_i.$$

The inner-product induced norm of a vector $\mathbf{x}$ (the Euclidean norm) is

$$\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{d} x_i^2,$$

where the latter two equalities are for the Euclidean inner product. The Cauchy-Schwarz inequality bounds the inner-product by the induced norms,

$$\langle \mathbf{x}, \mathbf{z} \rangle^2 \leq \|\mathbf{x}\|^2 \|\mathbf{z}\|^2,$$

with equality if and only if $\mathbf{x}$ and $\mathbf{z}$ are linearly dependent. The special case of the Cauchy-Schwarz inequality for the Euclidean inner product was given earlier in the chapter, $(\mathbf{x} \bullet \mathbf{z}) = (\sum_{i=1}^{d} x_i z_i)^2 \leq \sum_{i=1}^{d} x_i^2 \sum_{j=1}^{d} z_j^2$. The Pythagorean theorem applies to the square of a sum of vectors, and can be obtained by expanding $(\mathbf{x} + \mathbf{y})^{\mathrm{T}}(\mathbf{x} + \mathbf{y})$ to obtain:

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\mathbf{x}^{\mathrm{T}}\mathbf{y}.$$

If $\mathbf{x}$ is orthogonal to $\mathbf{y}$, then $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, which is the familiar Pythagorean theorem from geometry where $\mathbf{x}+\mathbf{y}$ is the diagonal of the triangle with sides given by the vectors $\mathbf{x}$ and $\mathbf{y}$.

Associated with any vector norm is the spectral (or operator) matrix norm $\|A\|_2$ which measures how large a vector transformed by A can get.

$$\|A\|_2 = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|;$$

If U has orthonormal columns then $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for any $\mathbf{x}$. Analogous to the Euclidean norm of a vector is the Frobenius matrix norm $\|A\|_F$ which sums the squares of the entries:

$$\|A\|_F^2 = \text{trace}(AA^{\mathrm{T}}) = \text{trace}(A^{\mathrm{T}}A) = \sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2.$$

Exercise B.6

Using the SVD of A, $A = U\Gamma V^{\mathrm{T}}$, show that

$$\begin{aligned} \|A\|_2 &= \gamma_1 && \text{(top singular value);} \\ \|A\|_F^2 &= \sum_{i=1}^{\rho}\gamma_i^2 && \text{(sum of squared singular values).} \end{aligned}$$

Note that

$$\|A\|_{2,F} = \|A^{\mathrm{T}}\|_{2,F};$$
$$\|A\|_2 \le \|A\|_F \le \sqrt{\rho}\|A\|_2,$$

where $\rho = \text{rank}(A)$. The matrix norms satisfy the triangle inequality and a property known as submultiplicativity, which bounds the norm of a product,

$$\begin{aligned} \|A+B\|_{2,F} &\le \|A\|_{2,F} + \|B\|_{2,F}; \\ \|AB\|_2 &\le \|A\|_2\|B\|_2; \\ \|AB\|_F &\le \|A\|_2\|B\|_F. \end{aligned}$$

The generalized Pythagorean theorem states that if $A^{\mathrm{T}}B = 0$ or $AB^{\mathrm{T}} = 0$ then

$$\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2$$
$$\max\{\|A\|_2^2, \|B\|_2^2\} \le \|A+B\|_2^2 \le \|A\|_2^2 + \|B\|_2^2.$$

B.5.1 Linear Hyperplanes and Quadratic Forms

A linear scalar function has the form

$$\ell(\mathbf{x}) = w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x},$$

where $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$. A hyperplane in d dimensions is defined by all points $\mathbf{x}$ satisfying $\ell(\mathbf{x}) = 0$. For a generic point $\mathbf{x}$, its geometric distance to the hyperplane is given by

$$\text{distance}(\mathbf{x}, (w_0, \mathbf{w})) = \frac{|w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}|}{\|\mathbf{w}\|}.$$

The vector $\mathbf{w}$ is normal to the hyperplane. A quadratic form $q(\mathbf{x})$ is a scalar function

$$q(\mathbf{x}) = w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x} + \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathrm{Q}\mathbf{x}.$$

When Q is positive definite, the manifold $q(\mathbf{x}) = 0$ defines an ellipsoid in d-dimensions (recall that $\mathbf{x} \in \mathbb{R}^d$).

B.6 Vector and Matrix Calculus

Let $E(\mathbf{x})$ be a scalar function of a vector $\mathbf{x} \in \mathbb{R}^d$. The gradient $\nabla E(\mathbf{x})$ is a d-dimensional vector function of $\mathbf{x}$ whose components are the partial derivatives; the Hessian $\mathrm{H}_E(\mathbf{x})$ is a $d \times d$ matrix function of $\mathbf{x}$ whose entries are the second order partial derivatives:

$$\begin{aligned} [\nabla E(\mathbf{x})]_i &= \frac{\partial}{\partial x_i} E(\mathbf{x}); \\ [\mathrm{H}_E(\mathbf{x})]_{ij} &= \frac{\partial^2}{\partial x_i \partial x_j} E(\mathbf{x}). \end{aligned}$$

The gradient and Hessian of the linear and quadratic forms are:

$$\begin{array}{rll} \text{linear form:} & \nabla \ell(\mathbf{x}) = \mathbf{w}; & \mathrm{H}_\ell(\mathbf{x}) = 0 \\ \text{quadratic form:} & \nabla q(\mathbf{x}) = \mathbf{w} + \mathrm{Q}\mathbf{x}; & \mathrm{H}_q(\mathbf{x}) = \mathrm{Q} \end{array}$$

For the general quadratic term $\mathbf{x}^{\mathrm{T}}\mathrm{A}\mathbf{x}$, the gradient is

$$\nabla(\mathbf{x}^{\mathrm{T}}\mathrm{A}\mathbf{x}) = (\mathrm{A} + \mathrm{A}^{\mathrm{T}})\mathbf{x}.$$

A necessary and sufficient condition for $\mathbf{x}^*$ to be a local minimum of $E(\mathbf{x})$ is that the gradient at $\mathbf{x}^*$ is zero and the Hessian at $\mathbf{x}^*$ is positive definite. The Taylor expansion of $E(\mathbf{x})$ around $\mathbf{x}_0$ up to second order terms is

$$E(\mathbf{x}) = E(\mathbf{x}_0) + \nabla E(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}}\mathrm{H}_{\mathrm{E}}(\mathbf{x})(\mathbf{x} - \mathbf{x}_0) + \cdots .$$

If $\mathbf{x}_0$ is a local minimum, the gradient term vanishes and the Hessian term is positive.

B.6.1 Multidimensional Integration

If $\mathbf{z} = \mathbf{q}(\mathbf{x})$ is a vector function of $\mathbf{x}$, then the Jacobian matrix J contains the derivatives of the components of $\mathbf{z}$ with respect to the components of $\mathbf{x}$:

$$[\mathrm{J}]_{ij} = \frac{\partial z_j}{\partial x_i}.$$

The Jacobian is important when performing a change of variables from $\mathbf{x}$ to $\mathbf{z}$ in a multidimensional integral, because it relates the volume element in $\mathbf{x}$-space to the volume element in $\mathbf{z}$-space. Specifically, the volume elements are related by

$$d\mathbf{x} = \frac{1}{|\mathrm{J}|} d\mathbf{z}.$$

As an example of using the Jacobian to transform variables in an integral, consider the multidimensional Gaussian distribution

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix. We can evaluate the expectations

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \int d\mathbf{x}\ \mathbf{x} \cdot P(\mathbf{x}) \qquad \text{and} \\
\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \int d\mathbf{x}\ \mathbf{x}\mathbf{x}^{\mathrm{T}} \cdot P(\mathbf{x})
\end{aligned}$$

by making a change of variables. Specifically, let $\Sigma = \mathrm{U}\Lambda\mathrm{U}^{\mathrm{T}}$, and change variables to $\mathbf{z} = \Lambda^{-1/2}\mathrm{U}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$. The Jacobian is $\mathrm{J} = \Lambda^{-1/2}\mathrm{U}^{\mathrm{T}}$ with determinant $|\mathrm{J}| = |\Lambda|^{-1/2}$. The integrals for the expectations then transform to

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \int d\mathbf{z}\ (\mathrm{U}\Lambda^{1/2}\mathbf{z} + \boldsymbol{\mu}) \cdot \frac{e^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}}}{(2\pi)^{d/2}} \\
&= \boldsymbol{\mu} \\
\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \int d\mathbf{z}\ (\mathrm{U}\Lambda^{1/2}\mathbf{z}\mathbf{z}^{\mathrm{T}}\Lambda^{1/2}\mathrm{U}^{\mathrm{T}} + \mathrm{U}\Lambda^{1/2}\mathbf{z}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\mathbf{z}^{\mathrm{T}}\Lambda^{1/2}\mathrm{U}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}) \cdot \frac{e^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}}}{(2\pi)^{d/2}} \\
&= \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}.
\end{aligned}$$

Exercise B.7

Show that the expressions for the expectations above do indeed transform to the integrals claimed in the formulae above. Carry through the integrations (filling in the necessary steps using standard techniques) to obtain the final results that are claimed above.

B.6.2 Matrix Derivatives

We now consider derivatives of functions of a matrix X. Let $q(\mathrm{X})$ be a scalar function of a matrix X. The derivative $\partial q(\mathrm{X})/\partial \mathrm{X}$ is a matrix of the same size as X with entries

$$\left[\frac{\partial}{\partial \mathrm{X}} q(\mathrm{X})\right]_{ij} = \frac{\partial}{\partial \mathrm{X}_{ij}} q(\mathrm{X}).$$

Similarly if a matrix X is a function of a scalar z then the the derivative $\partial \mathrm{X}(z)/\partial z$ is a matrix of the same size as X obtained by taking the derivative of each entry of X:

$$\left[\frac{\partial}{\partial z}\mathrm{X}(z)\right]_{ij} = \frac{\partial}{\partial z}\mathrm{X}_{ij}(z).$$

The matrix derivatives of several interesting functions can be expressed in a convenient form:

Function	$\frac{\partial}{\partial \mathrm{X}}$
$\mathrm{trace}(\mathrm{AXB})$	$\mathrm{A^T B^T}$
$\mathrm{trace}(\mathrm{AXX^T B})$	$\mathrm{A^T B^T X + BAX}$
$\mathrm{trace}(\mathrm{X^T A X})$	$(\mathrm{A} + \mathrm{A^T})\mathrm{X}$
$\mathrm{trace}(\mathrm{X^{-1} A})$	$-\mathrm{X^{-1} A^T X^{-1}}$
$\mathrm{trace}(\theta(\mathrm{BX})\mathrm{A})$	$\mathrm{B^T}\left(\theta'(\mathrm{BX}) \otimes \mathrm{A^T}\right)$
$\mathrm{trace}(\mathrm{A}\theta(\mathrm{BX})^{\mathrm{T}}\theta(\mathrm{BX}))$	$\mathrm{B^T}\left(\theta'(\mathrm{BX}) \otimes [\theta(\mathrm{BX})(\mathrm{A} + \mathrm{A^T})]\right)$
$\mathbf{a}^{\mathrm{T}}\mathrm{X}\mathbf{b}$	$\mathbf{a}\mathbf{b}^{\mathrm{T}}$
$\lvert\mathrm{AXB}\rvert$	$\lvert\mathrm{AXB}\rvert(\mathrm{X^T})^{-1}$
$\ln\lvert\mathrm{X}\rvert$	$(\mathrm{X^T})^{-1}$

(In all cases, assume the matrix products are well defined and the argument to the trace is a square matrix. When A and B have the same size $\mathrm{A} \otimes \mathrm{B}$ denotes component-wise multiplication, $[\mathrm{A} \otimes \mathrm{B}]_{ij} = A_{ij}B_{ij}$. A function applied to a matrix denotes the application of the function to each element, $[\theta(\mathrm{A})]_{ij} = \theta(A_{ij})$; θ' is the function which is the functional derivative of θ.) We can also get the derivatives of matrix functions of a parameter z:

Function	$\frac{\partial}{\partial z}$
X^{-1}	$-\mathrm{X}^{-1}\frac{\partial \mathrm{X}}{\partial z}\mathrm{X}^{-1}$
AXB	$\mathrm{A}\frac{\partial \mathrm{X}}{\partial z}\mathrm{B}$
XY	$\mathrm{X}\frac{\partial \mathrm{Y}}{\partial z} + \frac{\partial \mathrm{X}}{\partial z}\mathrm{Y}$
$\ln\lvert\mathrm{X}\rvert$	$\mathrm{trace}\left(\mathrm{X}^{-1}\frac{\partial \mathrm{X}}{\partial z}\right)$

$$\mathrm{X} = \mathrm{X}(z);\ \mathrm{Y} = \mathrm{Y}(z)$$

